

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

日 本 国 特 許 庁

PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日

Date of Application:

1999年 1月29日

出 願 番 号

Application Number:

平成11年特許願第022915号

出 願 人

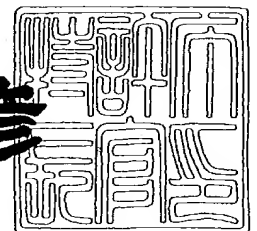
Applicant(s):

株式会社リコー

1999年10月22日

特許庁長官
Commissioner,
Patent Office

近 藤 隆 彦



出証番号 出証特平11-3072808

【書類名】 特許願

【整理番号】 9806050

【提出日】 平成11年 1月29日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 3/00

【発明の名称】 文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】 15

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 嶋田 敦夫

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 宮地 達生

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 剣持 栄治

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 山崎 真湖人

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 武谷 一寿

【発明者】

【住所又は居所】 東京都大田区中馬込 1 丁目 3 番 6 号 株式会社リコー内

【氏名】 長束 哲郎

【特許出願人】

【識別番号】 000006747

【氏名又は名称】 株式会社リコー

【代理人】

【識別番号】 100104190

【弁理士】

【氏名又は名称】 酒井 昭徳

【手数料の表示】

【予納台帳番号】 041759

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9810808

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項1】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項2】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、

前記表示手段に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示手段と、

前記クラスタ選択指示手段により選択されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項3】 前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを前記分類手段により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正手段と、

を備え、

前記分類手段は、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項2に記載の文書分類装置。

【請求項4】 前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項2に記載の文書分類装置。

【請求項5】 前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項3に記載の文書分類装置。

【請求項6】 前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、

前記表示手段は、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする請求項2または4に記載の文書分類装置。

【請求項7】 前記分類体系記憶手段は、前記選択指示手段により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報を分類体系の構成要素として記憶することを特徴とする請求項2～6に記載の文書分類装置。

【請求項8】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項 9】 文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、

前記表示工程に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示工程と、

前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項 10】 前記ベクトル生成工程により生成された文書特徴ベクトルを前記分類工程により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正工程と、

を含み、

前記分類工程は、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項 9 に記載の文書分類方法。

【請求項 11】 前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を

用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項 9 に記載の文書分類方法。

【請求項 12】 前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項 10 に記載の文書分類方法。

【請求項 13】 前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、

前記表示工程は、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする請求項 9 または 11 に記載の文書分類方法。

【請求項 14】 前記分類体系生成工程は、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする請求項 9 ～ 13 に記載の文書分類方法。

【請求項 15】 前記請求項 8 ～ 14 のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

この発明は、入力された複数の文書をその文書の内容に基づいて分類をおこなう、特に文書分類の際に算出される分類カテゴリー（体系）を精錬化する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを

記録したコンピュータ読み取り可能な記録媒体に関する。

【0002】

【従来の技術】

近年、インターネット等のネットワーク技術の普及により国内外の大量の電子化文書へのアクセスが可能になり、それに比例して業務上電子的に蓄積される情報の量も飛躍的に拡大した。その中で収集した大量の文書情報を意味あるカテゴリー（体系）に分類する等の知的作業の必要性が高まってきている。

【0003】

これらの大量の文書情報を意味的に分類するという作業の目的は、以下のようなものである。まず第1に、検索容易性の向上が考えられる。これは、膨大な文書群を分類名称（内容名）を手がかりに検索できるので検索が比較的容易である。

【0004】

第2に、情報群全体の把握が考えられる。これは、文書群全体がどのような内容（個々の分類）で構成されているかを把握する。しかし、大量の文書情報を操作者が手動で分類する場合、正確な分類をすることはできるが、分類に係る人的・時間的コストが膨大なものになるため、近年の文書分類の蓄積量の膨大さから、文書情報の自動分類装置が提案されるようになってきた。

【0005】

文書自動分類装置の従来技術としては、たとえば、特開平7-36897号公報に記載されているように、文書を、単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルを群分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【0006】

また、「Projections for Efficient Document Clustering（著者名：Hinrich Schutze and Craing Silverstein，学会名：ACM，論文名：Proceedings of SIGIR，ページ：74-81，発行年：1997）」においては、潜在的意味空間において文書分類を実施しているも

のがある。そのほかの方法としては、確率論的アプローチを用いる方法等が考えられる。

【0007】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、本質的には単語で構成される多次元空間に布置した文書を統計的な分類をする方法であるため、分類結果は単語のいわゆる振る舞いという観点から統計的に求められたものにすぎず、分類の結果、算出される各クラス（分類された個々の文書の部分集合）が操作者（利用者）に理解不能な場合がある。

【0008】

また、どのような分類結果が得られるかは分類対象のクラスタそのものに依存するため、最適な分類結果について定義することが困難であるという問題点があった。特に、上記情報群全体の把握に関し、多様な操作者の意図により要求される分類も異なるため、一度の分類作業で、操作者の所望する結果を得ることが困難であるという問題点があった。

【0009】

このように、文書分類の結果は、多くのいわゆるノイズを含んだものであると解釈することができ、その一部についてのみが操作者にとって有益な場合が多いという問題点があった。

【0010】

この発明は、上述した従来例による問題点を解消するため、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【0011】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項1の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解

析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【0012】

この請求項1の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【0013】

また、請求項2の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、前記表示手段に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示手段と、前記クラスタ選択指示手段により選択されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【0014】

この請求項2の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0015】

また、請求項3の発明に係る文書分類装置は、請求項2の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを前記分類手段により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正手段と、を備え、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【0016】

この請求項3の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0017】

また、請求項4の発明に係る文書分類装置は、請求項2の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0018】

この請求項4の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0019】

また、請求項5の発明に係る文書分類装置は、請求項3の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選

択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0020】

この請求項5の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0021】

また、請求項6の発明に係る文書分類装置は、請求項2または4の発明において、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、前記表示手段が、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする。

【0022】

この請求項6の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【0023】

また、請求項7の発明に係る文書分類装置は、請求項2～6の発明において、前記分類体系記憶手段が、前記選択指示手段により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報を分類体系の構成要素として記憶することを特徴とする。

【0024】

この請求項7の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができる。

【0025】

また、請求項 8 の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【0026】

この請求項 8 の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【0027】

また、請求項 9 の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、前記表示工程に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示工程と、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【0028】

この請求項 9 の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0029】

また、請求項 10 の発明に係る文書分類方法は、請求項 9 の発明において、前記ベクトル生成工程により生成された文書特徴ベクトルを前記分類工程により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正工程と、を含み、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【0030】

この請求項 10 の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0031】

また、請求項 11 の発明に係る文書分類方法は、請求項 9 の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0032】

この請求項 11 の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0033】

また、請求項 12 の発明に係る文書分類方法は、請求項 10 の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書

表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【0034】

この請求項12の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0035】

また、請求項13の発明に係る文書分類方法は、請求項9または11の発明において、前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、前記表示工程が、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする。

【0036】

この請求項13の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【0037】

また、請求項14の発明に係る文書分類方法は、請求項9～13の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする。

【0038】

この請求項14の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができる。

【0039】

また、請求項15の発明に係る記憶媒体は、請求項8～14に記載された方法

をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項 8～14 の動作をコンピュータによって実現することが可能である。

【0040】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【0041】

なお、以下説明する実施の形態においては、上記のように多くのノイズを含んだものであるとの解釈に基づいて、一回の文書集合からの話題（内容）抽出と位置づけ、文書分類のためのパラメータ（対象文書集合やクラス数、類似度測度、ストップワード等）を変化させながら複数化の分類を実行させ、その結果を保持・統合する手段を設けることで、任意の文書集合にどのような内容が含まれるかを漸次的に収集するものである。

【0042】

（実施の形態 1）

まず、この発明の実施の形態 1 による文書分類装置を構成する情報処理システム全体のハードウェア構成を説明する。図 1 は、本実施の形態による文書分類装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【0043】

図 1 において、実施の形態 1 による文書分類装置を構成する情報処理システムは、サーバー／クライアント方式で構成されている。すなわち、サーバー 101 と複数のクライアント 102 がネットワーク 103 によって接続されている。クライアント 102 は、分類データ等の加工データの生成、サーバー 101 への指示、分類結果等の加工処理結果の表示などをおこなう。一方、クライアント 102 からの指示にしたがって、サーバー 101 は文書（テキスト）分類等の加工処理を膨大な数値演算によりおこない、その処理の結果をクライアント 102 へ送る。

【0044】

分類処理の場合、より具体的には、サーバー101においては、テキスト分類処理（前処理、クラスタリング処理）がおこなわれ、クライアント102においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。サーバー101における処理は、上述のように、「前処理」と「分類処理」の二つに分かれており、その処理はデータによっては非常に負荷が大きくなる。したがって、サーバー101は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。

【0045】

また、サーバー101とクライアント102との間のデータのやりとりはファイル共有という方法を用いる。すなわち、分類処理等の加工処理に用いるファイルをサーバー101上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント102からはサーバー101の共有フォルダをネットワーク共有して利用することが可能である。

【0046】

つぎに、サーバー101およびクライアント102のハードウェア構成について説明する。図2は、実施の形態1による文書分類装置を構成する情報処理システムにおけるサーバー101をハードウェア的に示す説明図である。サーバー101は、たとえばワークステーション（WS）等が用いられる。

【0047】

図2において、201はサーバー101全体を制御するCPUを、202はブートプログラム等を記憶したROMを、203はCPU201のワークエリアとして使用されるRAM203を、204は通信回線205を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス（I/F）を、206はデータを記憶するディスク装置を示している。200は上記各部を結合させるためのバスを示している。

【0048】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ208や

、データを入力するためのキーボード209およびマウス210等が同様に接続されている。さらに、ディスク装置206には、クライアント102との間のデータのやりとりをするための共有フォルダ207が設けられている。

【0049】

また、図3は、実施の形態1による文書分類装置を構成する情報処理システムにおけるクライアント102をハードウェア的に示す説明図である。クライアント102は、たとえばパーソナルコンピュータ（PC）等が用いられる。

【0050】

図3において、301はシステム全体を制御するCPUを、302はブートプログラム等を記憶したROMを、303はCPU301のワークエリアとして使用されるRAMを、304はCPU301の制御にしたがってHD（ハードディスク）305に対するデータのリード／ライトを制御するHDD（ハードディスクドライブ）を、305はHDD304の制御で書き込まれたデータを記憶するHDを、306はCPU301の制御にしたがってFD（フロッピーディスク）307に対するデータのリード／ライトを制御するFDD（フロッピーディスクドライブ）を、307はFDD306の制御で書き込まれたデータを記憶する着脱自在のFDを、308はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【0051】

また、309は通信回線310を介してネットワーク103に接続され、そのネットワーク103と内部のインターフェイスを司るインターフェイス（I/F）を、311は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、312はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、313はOCR（Optical Character Reader）機能を備えた画像を光学的に読み取るスキャナを、314は分類結果を含むデータの内容等を印刷するプリンタを、315は上記各部を結合するためのバスをそれぞれ示している。また、HD305にはワープロソフト、表計算ソフト等のアプリケーションソフト316が記憶されている。

【0052】

つぎに、実施の形態1による文書分類装置の機能的構成について説明する。図4は、実施の形態1による文書分類装置の構成を機能的に示すブロック図である。

【0053】

図4のブロック図において、文書分類装置は、入力部401と、言語解析部402と、ベクトル生成部403と、分類部404と、分類パラメータ指示部405と、分類結果記憶部406と、クラスタ特徴表示部407と、クラスタ特徴算出部408と、分類体系記憶部409と、クラスタ選択指示部410と、分類体系閲覧操作部411と、を含む構成である。

【0054】

入力部401、言語解析部402、ベクトル生成部403、分類部404、分類パラメータ指示部405、分類結果記憶部406、クラスタ特徴表示部407、クラスタ特徴算出部408、分類体系記憶部409、クラスタ選択指示部410、分類体系閲覧操作部411は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0055】

ここで、入力部401は、文書データを入力するものであり、たとえば、キーボード209または311、スキャナ313、OCR機能を備えたスキャナ313、またはネットワーク103を経由して文書や文書群を得ることができるI/F204または309等である。

【0056】

また、入力部401は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。

【0057】

また、言語解析部 402 は、入力部 401 により入力された文書データを解析して言語解析情報を得るものであり、ベクトル生成部 403 は、言語解析部 402 により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するものである。

【0058】

また、分類部 404 は、ベクトル生成部 403 により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成するものであり、分類パラメータ指示部 405 は、分類パラメータを指示するものであり、たとえば、キーボード 209 または 311、マウス 210 または 312、またはネットワーク 103 を経由して指示情報を得ることができる I/F 204 または 309 等である。

【0059】

また、分類結果記憶部 406 は、分類部 404 により分類された結果、すなわち、分類された文書の部分集合に関する情報を記憶するものである。また、クラスタ特徴表示部 407 は、クラスタ特徴算出部 408 により算出されたクラスタ特徴を表示する。

【0060】

クラスタ特徴算出部 408 は、分類部 404 により生成された文書の部分集合の特徴であるクラスタ特徴を算出するものである。また、分類体系記憶部 409 は、クラスタ選択指示部 410 により選択されたクラスタ特徴を分類体系の構成要素として記憶するものである。

【0061】

クラスタ選択指示部 410 は、表示部 407 に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するものである。また、分類体系閲覧操作部 411 は、分類体系記憶部 409 に記憶されたデータを閲覧したい場合に、その閲覧の操作をおこなうものである。

【0062】

つぎに、文書集合に含まれる話題（内容）を抽出することが重要となる好適な例を、アンケート調査等により得られた自由記述回答の分析場面を想定し、その

具体例を用いて説明する。

【0063】

近年、たとえば、インターネット等を介して短期間に数千～数万件の自由記述回答を回収することが可能であり、このような機能を用いて大量のテキスト情報の収集をおこなうことができる。

【0064】

アンケート調査により得られた大量のテキスト情報の収集の例として、「オフィスのネットワーク化による無駄を挙げてください」という質問に対して文書で答えた一つの回答記述を文書とすると、文書集合（クラスタ）は1件ごとの回答の集合ということになる。

【0065】

ここで、操作者（アンケートの分析者）は、そのニーズの一つとして、意見集合（文書集合）にどのような種類の意見（話題）が含まれており、意見の概略を把握したい場合がある。このようなニーズを満たすべく、話題の抽出を類似する意見のまとまり（分類）により実現し、アンケート結果にどのような種類の意見が含まれているかを抽出する。

【0066】

文書分類は、典型的には大きく分けてつぎの3段階のステップから構成される。第1ステップでは、入力部401により入力された各文書（意見）について、言語解析部402が、各文書に含まれる単語（あるいは、特定の連続する文字列）を抽出する。この際、たとえば、形態素形跡等の言語解析アルゴリズムが用いられる。

【0067】

第2ステップでは、抽出された単語を列とし、各文書を行とし、要素を単語の出現頻度とした「単語」×「文書」の行列が生成される。なお、一般的な形態素解析機能と構文解析機能を有する言語解析ツールを用いると単語抽出のほかに、単語の品詞情報、複合語（フレーズ）、構文情報等の同時に取得することができ、こうした情報を上記単語×文書の行列を生成する際、考慮することができる。

【0068】

ベクトル生成部403は、この「単語」×「文書」の行列に基づいて単語で構成される多次元空間内に各文書をベクトル表現する。これには、以下の方法があり、本実施の形態においては、すべての方法を実装している。

【0069】

- (1) 行列の列成分をそのまま利用する方法、
 - (2) 各文書の長さ（文字の数やページ数等）や分類対象全体の文書集合内での各単語の出現頻度を考慮して値の重み付けをする方法、
 - (3) 上記行列から文書間の内積行列を算出し、これに特異値分解（たとえば、因子分析や主成分分析、数量化理論第3類等を利用しておこなわれる）を適用して潜在的意味空間を構成する方法、
- 等である。

【0070】

また、「Representating Documents Using an Explicit Model of Their Similarities（著者名：Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew, 論文名：Journal of the American Society for Information Science, 学会名：the American Society for Information Science, ページ：254-271, Vol. 46 No. 4, 発行年：1995）」

においては、上記潜在的意味空間への変換手法を一般化し、文書間の内積行列に、文書が有するほかの文書への参照情報から生成される共参照情報などを付加した行列を用いて、これらの類似性を反映する空間へ文書や単語を射影するための表現空間変換関数を導出しているものもあり、この方法も利用することができる。

【0071】

第3ステップでは、分類部404が、文書特徴ベクトルの類似度を用いて文書を分類する。具体的には分類対象データに対してカイ自乗法の手法、判別分析の方法、クラスタリングの方法等を適用することにより分類が実行される。

【0072】

また、類似度としては、内積や余弦、ユークリッド距離、マハラノビスの距離等が考えられ、本実施の形態においては、いずれの方法を用いてもよい。

【0073】

また、クラスタリングのアルゴリズムに関しても様々なものが公知になっている。クラスタリングは大別して階層型クラスタリングと非階層型クラスタリングが考えられるが、本実施の形態においては、いずれの方法を用いてもよい。

【0074】

また、分類パラメータ指示部405は、分類部404が文書特徴ベクトル进行分类するための分類パラメータを指示する。分類部404は、分類パラメータ指示部405により指示された分類パラメータにしたがって内部に保持される文書特徴ベクトル进行分类する。

【0075】

このようにして、第1ステップ～第3ステップの各処理を実行することにより第1回目の文書分類が終了すると、分類結果は分類結果記憶部406により保持される。

【0076】

引き続き、クラスタ特徴算出部408が、分類結果がどのようなクラスタを得ることができたのかを示す特徴、すなわちクラス特徴を算出する。典型的には各クラスタに所属する文書、あるいはその文書の一部を算出するが、その際、クラスタの重心との類似度に基づいて文書をソーティングして出力する。

【0077】

そのほか、クラスタ内で最頻の単語、クラスタに所属する文書数、クラスタ内の文書のばらつきの程度をあらわすクラスタ内の標準偏差のような数値をクラスタの特徴を表現するものとして算出する。

【0078】

これらのクラスタの特徴情報は、操作者に対して出力（表示）されたクラスタがどのようなもの（どのような特徴を有するもの）かを把握させるために算出されるものであり、操作者に対してクラスタの特徴を示すものであれば、上記の内

容（特徴）以外のものであってもよい。

【0079】

また、クラスタ特徴算出部408は、上記のようにクラスタの特徴を示すもの以外に、クラスタ間の関係を示す情報も算出する。階層型クラスタリングの場合は、その上位あるいは下位のクラスタを、非階層型クラスタリングの場合は、クラスタ重心間の類似度に基づく近接のクラスタを算出する。

【0080】

つぎに、クラスタ特徴表示部407によるクラスタ特徴の表示およびクラスタ選択の内容について説明する。図5は、実施の形態1による文書分類装置のクラスタ特徴表示部407の表示の一例を示す説明図である。

【0081】

図5において、クラスタ単位で操作者ができるようになっており、各クラスタは「クラスタID」欄501、「メンバー数」欄502、「頻度の高い単語」欄503、「文書内容」欄504、「重心との類似度」欄505等の項目から構成される。

【0082】

「クラスタID」欄501には、クラスタのIDを示す番号が通し番号で付与され、表示される。「メンバー数」欄502はクラスタに所属する文書あるいは文書の一部の数が算出され、表示される。その中で頻度の高い単語が抽出され「頻度の高い単語」欄503に表示される。「文書内容」欄504には文書の内容が表示され、「重心との類似度」欄505には、数値化された重心との類似度が表示される。これにより、操作者の理解容易性が向上する。

【0083】

操作者は、表示された情報（特徴量）に基づいてクラスタについてその特徴を把握することができる。ここで、内容（特徴）が理解可能なクラスタが一つでもあれば、クラスタ選択指示部410によりクラスタを選択することができる。

【0084】

より具体的には、マウス210または312等によって、表示されているクラスタの所定の位置、たとえば、「クラスタID」欄501へカーソル510を移

動させ、その位置でクリックすることにより、当該クラスタIDのクラスタ全体を選択することができる。

【0085】

図5においては、「クラスタID」欄501がクリックされ、これにより、クラスタ全体が反転表示しており、当該クラスタ（クラスタID「1」）が選択されたことを示している。

【0086】

また、操作者は、内容が理解可能であるクラスタが存在しない場合は、分類パラメータ指示部405により分類パラメータの再設定をおこない、再度分類実行をおこなうことができる。

【0087】

クラスタ選択指示部410により選択されたクラスタIDに関するデータは分類体系記憶部409へ送信される。分類体系記憶部409は、このクラスタIDに関するデータに基づいてクラスタ特徴算出部408からクラスタに関する上記特徴量を検索し記憶する。

【0088】

また、分類体系記憶部409は、同様に、分類結果記憶部406から分類結果を検索し記憶する。さらに、分類体系記憶部409は、操作者により入力されたクラスタに関するコメント（たとえば、「ネットワークの維持費が高い」等）の情報を併せて記憶することもできる。このように、操作者が作成した情報を分類体系の構成要素として記憶することにより、分類体系の利用価値がより向上する。

【0089】

なお、分類体系記憶部409により記憶されたデータは、別途閲覧操作のインターフェイスを設けることにより、選択・保持するクラスタの内容の閲覧や、クラスタ間の意味的な関連を手動であるいは、保持されているクラスタ重心間の類似度等を用いて自動で、構造化・体系化することができる。

【0090】

つぎに、実施の形態1の文書分類装置の一連の処理の手順について説明する。

図6は、実施の形態1による文書分類装置の一連の処理の手順を示すフローチャートである。図6のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS601）。

【0091】

つぎに、入力された文書の言語が解析され（ステップS602）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成される（ステップS603）。

【0092】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS604肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS605）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS606）。

【0093】

つぎに、分類されたクラスタの特徴を算出し（ステップS607）、算出された結果を表示する（ステップS608）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS609）、選択されなかった場合（ステップS609否定）は、ステップS604へ移行し、再度分類パラメータの指示があるのを待つ（ステップS604）。

【0094】

一方、ステップS609において、クラスタが選択された場合（ステップS609肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS610）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。これにより、一連の処理を終了する。

【0095】

以上説明したように、実施の形態1による文書分類装置によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することができる。

【0096】

したがって、分類部404によりクラスタを得ることができるとともに、クラスタ特徴算出部408・分類体系記憶部409により、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【0097】

また、クラスタ選択指示部410により選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【0098】

(実施の形態2)

さて、上述した実施の形態1に加えて、以下に説明する実施の形態2のように、さらにベクトル記憶部と、ベクトル修正部とを含む構成とするようにしてもよい。

【0099】

実施の形態2による文書分類装置を構成する情報処理システムは、図1に示したように実施の形態1と同様であるので、その説明は省略する。また、サーバー101およびクライアント102のハードウェア構成についても、図2・図3に示したように実施の形態1と同様であるので、その説明は省略する。

【0100】

つぎに、実施の形態2による文書分類装置の機能的構成について説明する。図7は、この発明の実施の形態2による文書分類装置の構成を機能的に示すブロック図である。図7において、実施の形態1の図4と同一のものに関しては同じ符号を付して、その説明を省略する。

【0101】

図7のブロック図において、文書分類装置は、入力部401、言語解析部402、ベクトル生成部403、分類部404、分類パラメータ指示部405、分類結果記憶部406、クラスタ特徴表示部407、クラスタ特徴算出部408、分類体系記憶部409、クラスタ選択指示部410、分類体系閲覧操作部411の

ほか、ベクトル記憶部 701 と、ベクトル修正部 702 とを含む構成である。

【0102】

ベクトル記憶部 701 は、ベクトル生成部 403 により生成された文書特徴ベクトルを記憶するものである。また、ベクトル修正部 702 は、文書特徴ベクトル記憶部 701 により記憶された文書特徴ベクトルを分類部 404 により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するものである。

【0103】

また、分類部 404 は、ベクトル修正部 702 により修正された文書特徴ベクトルに基づいて文書を分類する。

【0104】

なお、ベクトル記憶部 701、ベクトル修正部 702 は、ROM 202 または 302、RAM 203 または 303、あるいはディスク装置 306 またはハードディスク 316 等の記録媒体に記録されたプログラムに記載された命令にしたがって CPU 201 または 301 等が命令処理を実行することにより、各部の機能を実現する。

【0105】

ベクトル生成部 403 において生成された文書特徴ベクトル（列ベクトル）・単語（単語特徴）ベクトル（行ベクトル）はベクトル記憶部 701 によって記憶される。これは、次回以降の分類実行の際に利用する文書特徴ベクトルを確保するためである。

【0106】

ベクトル修正部 702 は、クラスタ選択指示部 410 により選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除き、次回以降もこれらの文書が除かれるよう削除する。削除された文書特徴ベクトルはベクトル記憶部 701 により記憶される。

【0107】

この結果、ベクトル記憶部 701 に記憶されているベクトルデータのうち、選択されたクラスタに所属する文書（もしくは操作者に指定されたその一部）列ベ

クトルを除いたものが次回以降の分類実行の際に利用されるデータとなる。

【0108】

つぎに、実施の形態2の文書分類装置の一連の処理の手順について説明する。
図8は、実施の形態2による文書分類装置の一連の処理の手順を示すフローチャートである。図2のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS801）。

【0109】

つぎに、入力された文書の言語が解析され（ステップS802）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップS803）、生成された文書特徴ベクトルが記憶される（ステップS804）。

【0110】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS805肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS806）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS807）。

【0111】

つぎに、分類されたクラスタの特徴を算出し（ステップS808）、算出された結果を表示する（ステップS809）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS810）、選択されなかった場合（ステップS810否定）は、ステップS805へ移行し、再度分類パラメータの指示があるのを待つ（ステップS805）。

【0112】

一方、ステップS810において、クラスタが選択された場合（ステップS810肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS811）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップS812）。

【0113】

ステップ S 8 1 2 において、繰り返して処理をおこなう旨の指示があった場合（ステップ S 8 1 2 肯定）は、選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップ S 8 1 3）。その後、ステップ S 8 0 5 へ移行し、以後、ステップ S 8 0 5 ～ S 8 1 3 の各処理を繰り返しおこなう。

【0114】

一方、ステップ S 8 1 2 において、繰り返して処理をおこなう旨の指示がない場合（ステップ S 8 1 2 否定）は、これにより、一連の処理を終了する。

【0115】

以上説明したように、実施の形態 2 による文書分類装置によれば、ベクトル修正部 7 0 1 により、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【0116】

（実施の形態 3）

さて、上述した実施の形態 2 においては、ベクトル記憶部およびベクトル修正部とを含む構成であったが、以下に説明する実施の形態 3 のように、ベクトル修正部に代わりに、文書表現空間修正部を含む構成とするようにしてもよい。

【0117】

実施の形態 3 による文書分類装置を構成する情報処理システムは、図 1 に示したように実施の形態 1 と同様であるので、その説明は省略する。また、サーバー 1 0 1 およびクライアント 1 0 2 のハードウェア構成についても、図 2・図 3 に示したように実施の形態 1 と同様であるので、その説明は省略する。

【0118】

つぎに、実施の形態 3 による文書分類装置の機能的構成について説明する。図 9 は、この発明の実施の形態 3 による文書分類装置の構成を機能的に示すブロック図である。図 9 において、実施の形態 1 の図 4 と同一のものに関しては同じ符号を付して、その説明を省略する。

【0119】

図 9 のブロック図において、文書分類装置は、入力部 4 0 1、言語解析部 4 0

2、ベクトル生成部403、分類部404、分類パラメータ指示部405、分類結果記憶部406、クラスタ特徴表示部407、クラスタ特徴算出部408、分類体系記憶部409、クラスタ選択指示部410、分類体系閲覧操作部411のほか、ベクトル記憶部901と、文書表現空間修正部902とを含む構成である。

【0120】

ベクトル記憶部901は、ベクトル生成部403により生成された文書特徴ベクトルを記憶するものである。また、文書表現空間修正部902は、文書特徴ベクトル記憶部901により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部410により選択されたクラスタ特徴に基づいて修正するものである。

【0121】

また、分類部404は、文書表現空間修正部902により修正された文書表現空間を用いて、ベクトル生成部403により生成された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【0122】

なお、ベクトル記憶部901、文書表現空間修正部902は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、各部の機能を実現する。

【0123】

つぎに、文書表現空間修正部902の内容について説明する。実施の形態2におけるベクトル修正部702にあっては、既知になったクラスタの影響を排除するために文書特徴ベクトルを除去するが、文書特徴ベクトルを表現する多次元空間自体の変更はおこなわれない。

【0124】

したがって、前回の分類実行の結果、操作者により選択されたクラスタの形成特徴を次回の分類実行の際に排除したい場合は、文書ベクトルを表現する空間自

体の変更が必要となる。

【0125】

そこで、文書表現空間修正部902を備え、文書表現空間の修正をおこなうものである。ここで、文書表現空間の特徴次元を変更する例として、操作者により選択されたクラスタの重心と類似度の高い特徴次元の削除をおこなうことについて説明する。

【0126】

操作者により選択されたクラスタの重心はベクトルとして表現することができるので、このクラスタ重心ベクトルとベクトル記憶部901に記憶されている文書表現空間の各特徴次元との類似度を算出することにより、類似度の高い特徴次元を判別する。

【0127】

なお、類似の測度としては、余弦、内積、ユークリッド距離、マハラノビス距離等を用いる。また、判別に関してはある類似度以上を削除対象として採用するようなしきい値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理による判別を用いる。また、判別分析等も用いることができる。

【0128】

文書表現空間修正部902は、上述のような削除対象の特徴次元を算出して、特徴次元の削除をおこなう。特徴次元の削除は、ベクトル記憶部901に記憶されている「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することによりおこなう。文書表現空間修正部902により修正された文書ベクトルは、次回以降の分類のために、ベクトル記憶部901に記憶される。

【0129】

つぎに、実施の形態3の文書分類装置の一連の処理の手順について説明する。図10は、実施の形態3による文書分類装置の一連の処理の手順を示すフローチャートである。図10のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS1001）。

【0130】

つぎに、入力された文書の言語が解析され（ステップS1002）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップS1003）、生成された文書特徴ベクトルが記憶される（ステップS1004）。

【0131】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS1005肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS1006）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS1007）。

【0132】

つぎに、分類されたクラスタの特徴を算出し（ステップS1008）、算出された結果を表示する（ステップS1009）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS1010）、選択されなかった場合（ステップS1010否定）は、ステップS1005へ移行し、再度分類パラメータの指示があるのを待つ（ステップS1005）。

【0133】

一方、ステップS1010において、クラスタが選択された場合（ステップS1010肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS1011）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップS1012）。

【0134】

ステップS1012において、繰り返して処理をおこなう旨の指示があった場合（ステップS1012肯定）は、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップS1013）。その後、ステップS1005へ移行し、以後、ステップS1005～S1013の各処理を繰り返しおこなう。

【0135】

一方、ステップ S1012 において、繰り返して処理をおこなう旨の指示がなかった場合（ステップ S1012 否定）は、これにより、一連の処理を終了する。

【0136】

以上説明したように、実施の形態 3 による文書分類装置によれば、前回の分類実行の結果、文書表現空間修正部 902 により操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0137】

（実施の形態 4）

さて、上述した実施の形態 2 または実施の形態 3 においては、ベクトル修正部または文書表現空間修正部のいずれか一方のみを含む構成であったが、以下に説明する実施の形態 4 のように、ベクトル修正部および文書表現空間修正部の両方を含む構成とするようにしてもよい。

【0138】

実施の形態 4 による文書分類装置を構成する情報処理システムは、図 1 に示したように実施の形態 1 と同様であるので、その説明は省略する。また、サーバー 101 およびクライアント 102 のハードウェア構成についても、図 2・図 3 に示したように実施の形態 1 と同様であるので、その説明は省略する。

【0139】

つぎに、実施の形態 4 による文書分類装置の機能的構成について説明する。図 11 は、この発明の実施の形態 4 による文書分類装置の構成を機能的に示すブロック図である。図 11 において、実施の形態 1 の図 4 と同一のものに関しては同じ符号を付して、その説明を省略する。

【0140】

図 11 のブロック図において、文書分類装置は、入力部 401、言語解析部 402、ベクトル生成部 403、分類部 404、分類パラメータ指示部 405、分類結果記憶部 406、クラスタ特徴表示部 407、クラスタ特徴算出部 408、分類体系記憶部 409、クラスタ選択指示部 410、分類体系閲覧操作部 411

のほか、ベクトル記憶部 1101 と、ベクトル修正部 1102 と、文書表現空間修正部 1103 とを含む構成である。

【0141】

ベクトル記憶部 1101 は、ベクトル生成部 403 により生成された文書特徴ベクトルを記憶するものである。また、ベクトル修正部 1102 は、文書特徴ベクトル記憶部 1101 により記憶された文書特徴ベクトルを分類部 404 により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するものである。

【0142】

また、文書表現空間修正部 1103 は、文書特徴ベクトル記憶部 1101 により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部 410 により選択されたクラスタ特徴に基づいて修正するものである。

【0143】

また、分類部 404 は、文書表現空間修正部 1103 により修正された文書表現空間を用いて、ベクトル修正部 1102 により修正された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【0144】

なお、ベクトル記憶部 1101、ベクトル修正部 1102、文書表現空間修正部 1103 は、ROM202 または 302、RAM203 または 303、あるいはディスク装置 306 またはハードディスク 316 等の記録媒体に記録されたプログラムに記載された命令にしたがって CPU201 または 301 等が命令処理を実行することにより、各部の機能を実現する。

【0145】

つぎに、ベクトル修正部 1102 および文書表現空間修正部 1103 の内容について説明する。実施の形態 3 においては、選択されたクラスタに所属する文書は次回以降の分類実行の際にも使用される。

【0146】

実施の形態 4 では、ベクトル修正部 1102 および文書表現空間修正部 110

3の両方を具備することにより、選択されたクラスタに所属する文書を次回の分類実行の際に除去し、次回の分類実行の際には分類対象文書としないようにする。

【0147】

実施の形態3においては、話題抽出の側面を強調し、ある文書が複数の話題として分類される可能性を前提としており、たとえば、ネットワーク化に関する調査における「エンドユーザーがソフトウェアのインストール方法について聞いてくるのでシステム管理者としての仕事ができない」という回答について言えば、この意見は「ソフトウェアの操作方法理解に関する困難性」という話題として分類され得るし、「システム管理者の仕事の多忙さ」という話題で分類される可能性もある。

【0148】

実施の形態3においては、いずれにしても、「ソフトウェアの操作方法理解に関する困難性」というクラスタと「システム管理者の仕事の多忙さ」というクラスタの両方とも抽出したいというニーズに応えている。

【0149】

これとは反対に、操作者は、一度抽出した話題は既知であるので、次回の分類の際にはなるべく異なる分類結果が欲しいとするケースも考えられる。実施の形態4では、このような要求に応えるため、ベクトル修正部1102により、n回目の分類で選択されたクラスタに所属する文書のすべてまたはその一部を次回以降の分類を実行する際、分類対象から除去するものである。

【0150】

クラスタ選択指示部410により選択指示を受けたクラスタの所属文書はベクトル記憶部1101において列ベクトルの形式で記憶されているため、ベクトル修正部1102では劣ベクトルを除去することで、次回以降の分類実行用の分類対象文書集合を生成する。

【0151】

さらに、実施の形態3と同様に、選択されたクラスタにより文書表現空間修正部1103は、ベクトル記憶部1101に記憶されている行列から特徴次元を削

除する。

【0152】

つぎに、実施の形態4の文書分類装置の一連の処理の手順について説明する。
図12は、実施の形態4による文書分類装置の一連の処理の手順を示すフローチャートである。図12のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップS1201）。

【0153】

つぎに、入力された文書の言語が解析され（ステップS1202）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップS1203）、生成された文書特徴ベクトルが記憶される（ステップS1204）。

【0154】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップS1205肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップS1206）、その結果、すなわち、クラスタに関する情報を記憶する（ステップS1207）。

【0155】

つぎに、分類されたクラスタの特徴を算出し（ステップS1208）、算出された結果を表示する（ステップS1209）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップS1210）、選択されなかった場合（ステップS1210否定）は、ステップS1205へ移行し、再度分類パラメータの指示があるのを待つ（ステップS1205）。

【0156】

一方、ステップS1210において、クラスタが選択された場合（ステップS1210肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップS1211）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップS1212）。

【0157】

ステップ S 1 2 1 2 において、繰り返して処理をおこなう旨の指示があった場合（ステップ S 1 2 1 2 肯定）は、選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップ S 1 2 1 3）。

【0158】

ステップ S 1 2 1 3 に引き続き、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップ S 1 2 1 4）。その後、ステップ S 1 2 0 5 へ移行し、以後、ステップ S 1 2 0 5 ～ S 1 2 1 4 を繰り返しておこなう。

【0159】

一方、ステップ S 1 2 1 2 において、繰り返して処理をおこなう旨に指示がない場合（ステップ S 1 2 1 2 否定）は、これにより、一連の処理を終了する。

【0160】

以上説明したように、実施の形態 4 よる文書分類装置によれば、ベクトル修正部 1 1 0 2 が、既知になったクラスタの影響を排除し、かつ、文書表現空間修正部 1 1 0 3 が、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【0161】

（実施の形態 5）

さて、上述した実施の形態 1 または実施の形態 3 においては、繰り返し分類処理をおこなった場合に、ある文書が何度選択されたかその情報については考慮していなかったが以下に説明する実施の形態 5 のように、選択情報付与部を含む構成とし、選択情報をクラスタ特徴とともに表示するようにしてもよい。

【0162】

実施の形態 5 による文書分類装置を構成する情報処理システムは、図 1 に示したように実施の形態 1 と同様であるので、その説明は省略する。また、サーバー 1 0 1 およびクライアント 1 0 2 のハードウェア構成についても、図 2・図 3 に示したように実施の形態 1 と同様であるので、その説明は省略する。

【0163】

つぎに、実施の形態5による文書分類装置の機能的構成について説明する。図13は、この発明の実施の形態5による文書分類装置の構成を機能的に示すブロック図である。図13において、実施の形態3の図9と同一のものに関しては同じ符号を付して、その説明を省略する。

【0164】

図13のブロック図において、文書分類装置は、入力部401、言語解析部402、ベクトル生成部403、分類部404、分類パラメータ指示部405、分類結果記憶部406、クラスタ特徴表示部407、クラスタ特徴算出部408、分類体系記憶部409、クラスタ選択指示部410、分類体系閲覧操作部411、ベクトル記憶部901、文書表現空間修正部902のほか、選択情報付与部1301を含む構成である。

【0165】

選択情報付与部1301は、分類部404により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する。また、クラスタ特徴表示部407は、クラスタ特徴を表示するとともに、選択情報付与部1301により付与された選択情報を表示する。

【0166】

なお、選択情報付与部1301は、ROM202または302、RAM203または303、あるいはディスク装置306またはハードディスク316等の記録媒体に記録されたプログラムに記載された命令にしたがってCPU201または301等が命令処理を実行することにより、機能を実現する。

【0167】

つぎに、選択情報付与部1301の詳細な内容について説明する。アンケートの調査の例において、独自性の高いユニークな意見は貴重であることが経験的に知られている。これは、調査を企画する担当者が予想できなかった意見である場合が多いからである。

【0168】

そこで、操作者に選択されたクラスタに所属する文書を、次回以降の分類実行

の際に使用する場合において、クラスタ特徴表示部 407 で個々の文書を表示する際に、各文書が何回選択されたかを示すことで、多重に利用される文書の識別性を向上させ、かつ一度も選択されない文書の識別性も向上させることができる。

【0169】

図 14 は、実施の形態 5 による文書分類装置の分類結果記憶部 406 において設けられたテーブル 1400 を示す説明図である。図 14 において、文書 ID ごとにテーブル化されており、テーブル 1400 は、各文書がどのサイクルに分類実行の際に操作者に選択されたかを記録する。すなわち、選択された場合は選択情報として「1」を記録し、選択されなかった場合は選択情報として「0」を記録する。

【0170】

たとえば、4 回分類が実行された際、文書 ID の「1」、第 1 回目および第 2 回目の分類実行時に操作者に選択されたことを示し、第 3 回目、第 4 回目の分類実行時には選択されなかったことを示している。一方、文書 ID の「2」は、未だ一度も選択されておらず、操作者にとって未知の意見という可能性を示唆している。

【0171】

こうした情報に基づいて、クラスタ特徴表示部 407 が文書を操作者に表示する際、たとえば、選択された回数に応じて表示を変化させるようにするとよい。変化させる視覚的特性としては、たとえば文字や背景の色の濃度や彩度等が考えられる。

【0172】

また、直接的に数字やグラフ等で選択された回数を表現することもできる。いずれにしてもよく選択される文書と一度も選択されていない文書とを視覚的に識別できる表示形式であれば、上記のものに限らない。

【0173】

また、上記選択情報を分類体系閲覧操作部 411 の閲覧操作により閲覧できるようにしてもよい。

【0174】

つぎに、選択情報付与部1301の処理の内容について説明する。図15は、実施の形態5による文書分類装置の選択情報付与部1301の処理の手順を示すフローチャートである。図15のフローチャートにおいて、まず、分類処理がおこなわれ（ステップS1501）、それに引き続き、最初の文書が抽出される（ステップS1502）。

【0175】

抽出された文書が、ステップS1501における分類処理の際に選択されたか否かを判断する（ステップS1503）。ここで、選択された場合（ステップS1503肯定）は、選択情報としてデータ「1」を記録する（ステップS1504）。一方、選択されなかった場合（ステップS1503否定）は、選択情報としてデータ「0」を記録する（ステップS1505）。

【0176】

つぎに、すべての文書について処理が終了したか否かを判断する（ステップS1506）。ここで、すべての文書について処理が終了していない場合（ステップS1506否定）は、つぎに文書を抽出し（ステップS1507）、ステップS1503へ移行し、以後、ステップS1503～S1507を繰り返しおこなう。

【0177】

一方、ステップS1506において、すべての文書について処理が終了した場合（ステップS1506肯定）は、ステップS1501へ移行し、再度分類処理がおこなわれる（ステップS1501）。このようにして、分類処理がおこなわれる回数だけ、ステップS1501～S1507の各処理が繰り返しおこなわれる。

【0178】

以上説明したように、実施の形態5によれば、選択情報付与部1301が選択情報を付与し、その選択情報をクラスタ特徴表示部407が表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【0179】

なお、実施の形態1～5で説明した文書分類方法は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーション等のコンピュータで実行することにより実現される。このプログラムは、ハードディスク、フロッピーディスク、CD-ROM、MO、DVD等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、上記記録媒体を介して、インターネット等のネットワークを介して配布することができる。

【0180】

【発明の効果】

以上説明したように、請求項1の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系記憶手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶するので、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0181】

また、請求項2の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類

し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示し、クラスタ選択指示手段が、前記表示手段に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択し、分類体系記憶手段が、前記クラスタ選択指示手段により選択されたクラスタ特徴を分類体系の構成要素として記憶するので、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0182】

また、請求項3の発明によれば、請求項2の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、ベクトル修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを前記分類手段により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正し、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類するので、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0183】

また、請求項4の発明によれば、請求項2の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正し、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するので、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分

類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0184】

また、請求項5の発明によれば、請求項3の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正し、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するので、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0185】

また、請求項6の発明によれば、請求項2または4の発明において、選択情報付与手段が、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示手段が、前記クラスタ特徴を表示するとともに、選択情報付与手段により付与された選択情報を表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0186】

また、請求項7の発明によれば、請求項2～6の発明において、前記分類体系記憶手段が、前記選択指示手段により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは

一部および／または操作者が作成した情報を分類体系の構成要素として記憶するので、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【0187】

また、請求項8の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系生成工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0188】

また、請求項9の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示し、クラスタ選択指示工程が、前記表示工程に表示された複数のクラスタ特徴の中から所望のクラス

タ特徴を選択し、分類体系生成工程が、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0189】

また、請求項10の発明によれば、請求項9の発明において、ベクトル修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトルを前記分類工程により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正し、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類するので、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0190】

また、請求項11の発明によれば、請求項9の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するので、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【0191】

また、請求項12の発明によれば、請求項10の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似

度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するので、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【 0 1 9 2 】

また、請求項 1 3 の発明によれば、請求項 9 または 1 1 の発明において、選択情報付与工程が、前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示工程が、前記クラスタ特徴を表示するとともに、選択情報付与工程により付与された選択情報を表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【 0 1 9 3 】

また、請求項 1 4 の発明によれば、請求項 9 ～ 1 3 の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成するので、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【 0 1 9 4 】

また、請求項 1 5 の発明によれば、請求項 8 ～ 1 4 のいずれか一つに記載され

た方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項 8～14 の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図 1】

この発明の実施の形態 1 による文書分類装置を構成する情報処理システム全体のハードウェア構成を示す説明図である。

【図 2】

実施の形態 1 による文書分類装置を構成する情報処理システムにおけるサーバーをハードウェア的に示す説明図である。

【図 3】

実施の形態 1 による文書分類装置を構成する情報処理システムにおけるクライアントをハードウェア的に示す説明図である。

【図 4】

実施の形態 1 による文書分類装置の構成を機能的に示すブロック図である。

【図 5】

実施の形態 1 による文書分類装置のクラスタ特徴表示部の表示の一例を示す説明図である。

【図 6】

実施の形態 1 による文書分類装置の一連の処理の手順を示すフローチャートである。

【図 7】

この発明の実施の形態 2 による文書分類装置の構成を機能的に示すブロック図である。

【図 8】

実施の形態 2 による文書分類装置の一連の処理の手順を示すフローチャートである。

【図 9】

この発明の実施の形態 3 による文書分類装置の構成を機能的に示すブロック図

である。

【図 10】

実施の形態 3 による文書分類装置の一連の処理の手順を示すフローチャートである。

【図 11】

この発明の実施の形態 4 による文書分類装置の構成を機能的に示すブロック図である。

【図 12】

実施の形態 4 による文書分類装置の一連の処理の手順を示すフローチャートである。

【図 13】

この発明の実施の形態 5 による文書分類装置の構成を機能的に示すブロック図である。

【図 14】

実施の形態 5 による文書分類装置の分類結果記憶部において設けられたテーブルを示す説明図である。

【図 15】

実施の形態 5 による文書分類装置の選択情報付与部の処理の手順を示すフローチャートである。

【符号の説明】

101 サーバー

102 クライアント

103 ネットワーク

200 バス

201 CPU

202 ROM

203 RAM

204 I/F

205 通信回線

206 ディスク装置
 207 共有フォルダ
 208 ディスプレイ
 209 キーボード
 210 マウス
 301 CPU
 302 ROM
 303 RAM
 306 ハードディスク
 308 ディスプレイ
 309 I/F
 311 キーボード
 312 マウス
 313 スキャナ
 314 プリンタ
 401 入力部
 402 言語解析部
 403 ベクトル生成部
 404 分類部
 405 分類パラメータ指示部
 406 分類結果記憶部
 407 クラスタ特徴表示部
 408 クラスタ特徴算出部
 409 分類体系記憶部
 410 クラスタ選択指示部
 411 分類体系閲覧操作部
 510 カーソル
 701, 901, 1101 ベクトル記憶部
 702, 1102 ベクトル修正部

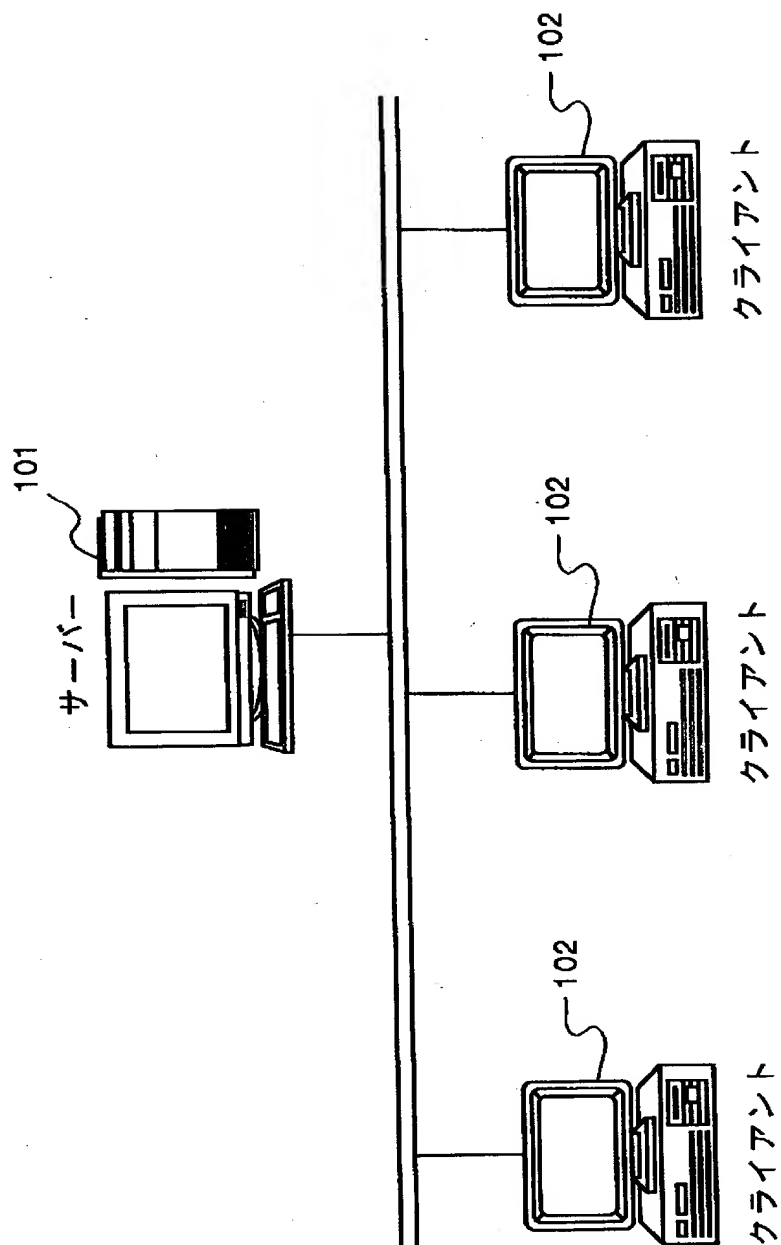
902, 1103 文書表現空間修正部

1301 選択情報付与部

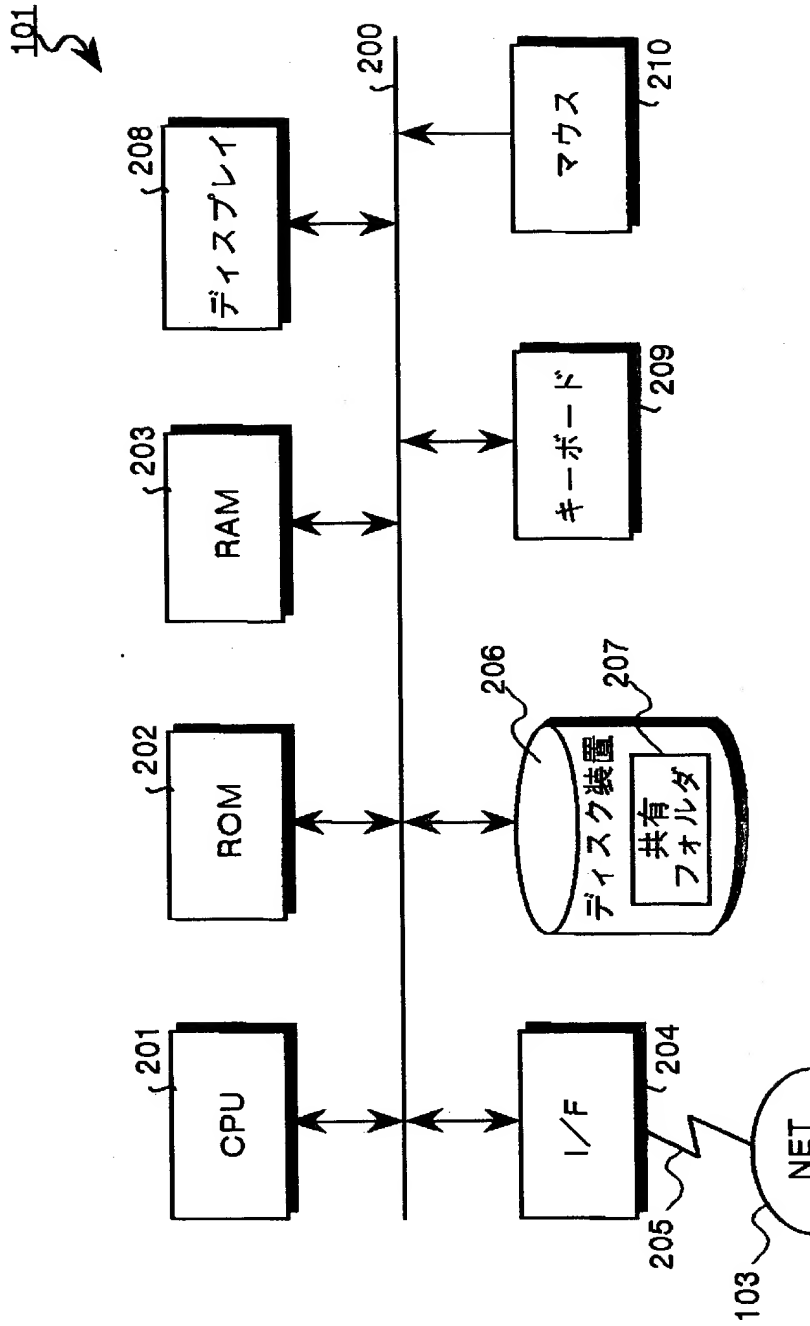
1400 テーブル

【書類名】 図面

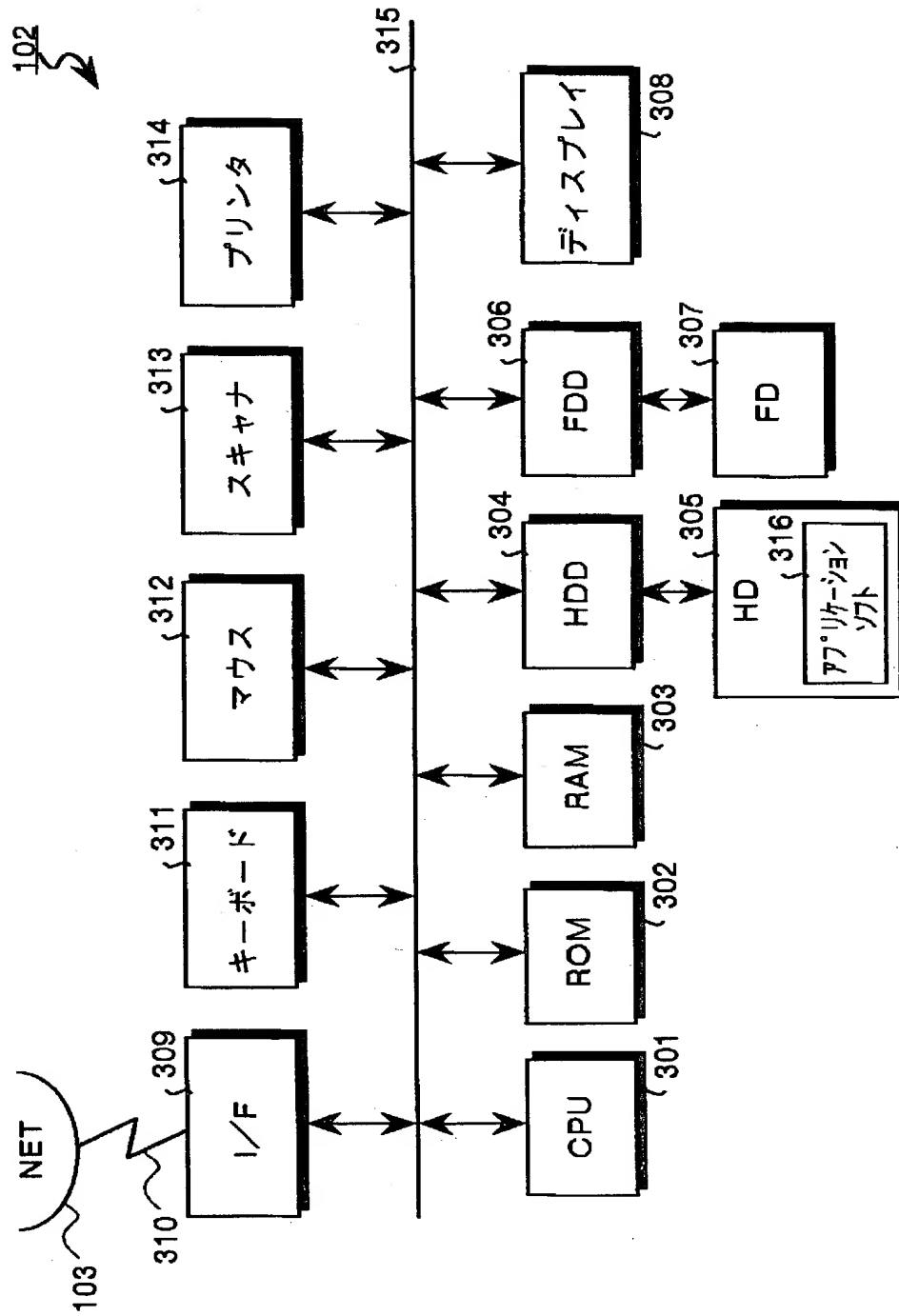
【図 1】



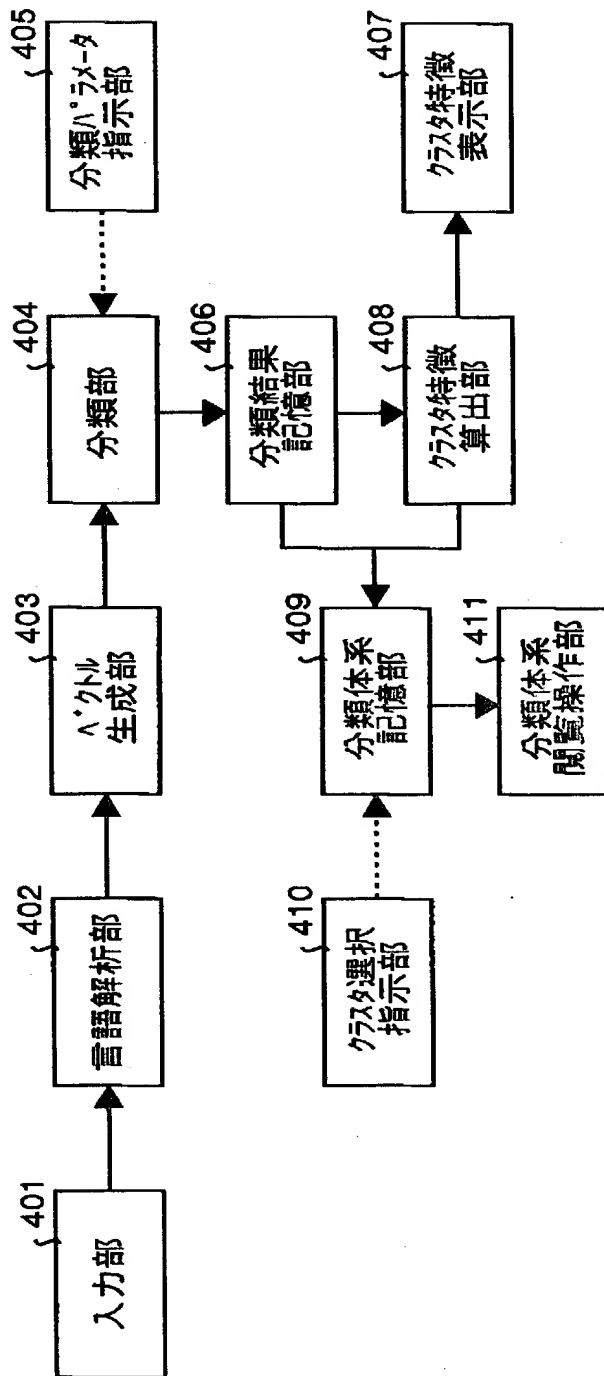
【図 2】



【図 3】



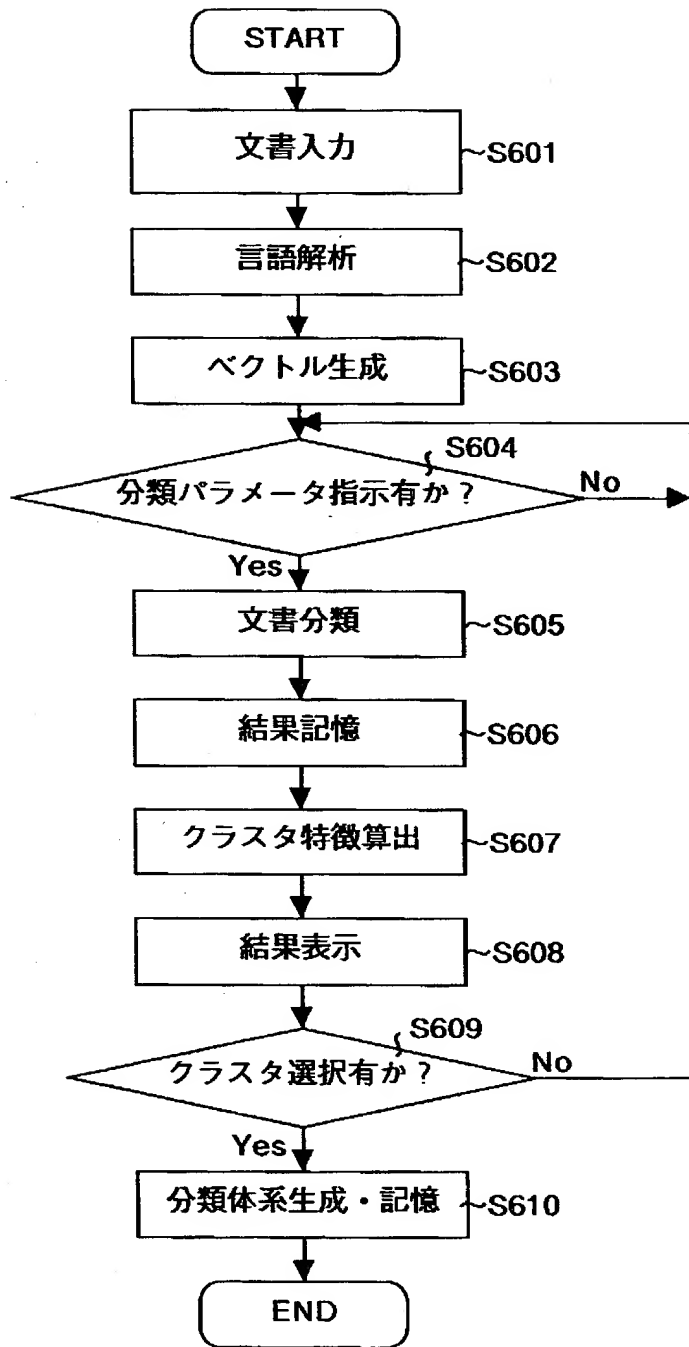
【図 4】



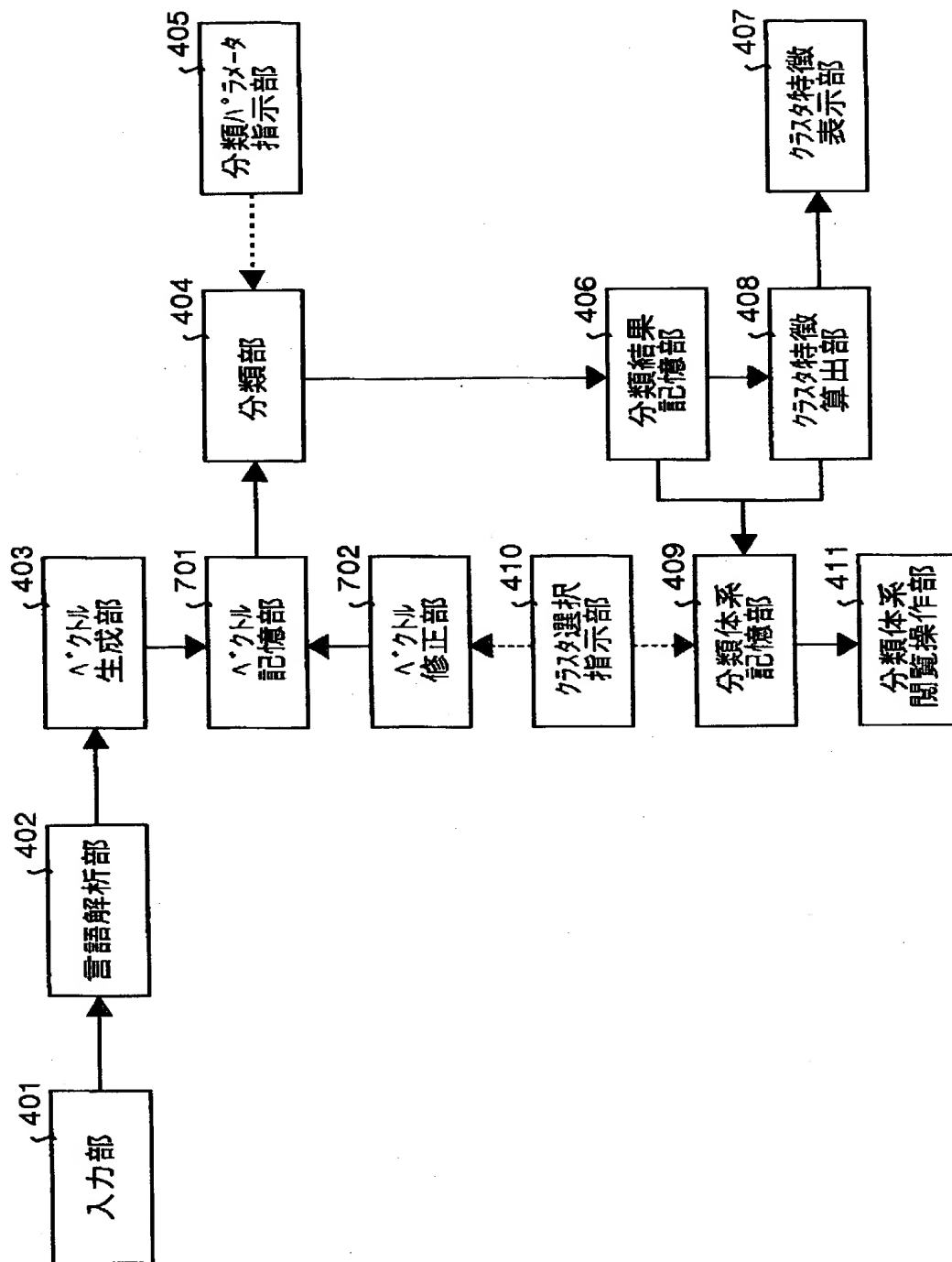
【図 5】

501		502	503	504	505
クラスID	メンバー数	頻度の高い単語	文書内容	重心との類似度	
1	248	管理者、多忙...	システム管理業務が多くて多忙だ	0.987	
510			システム管理が多忙でコスト削減できない	0.965	
			管理者が多忙だとシステムがダウンする	0.911	
			システムダウンで管理が大変	0.889	
			システムダウンで管理業務が多忙になる	0.876	
2					
N	1498	操作性、悪い...	ソフトの操作性が悪い	0.969	
			ソフトの操作性を覚えるの大変である	0.962	

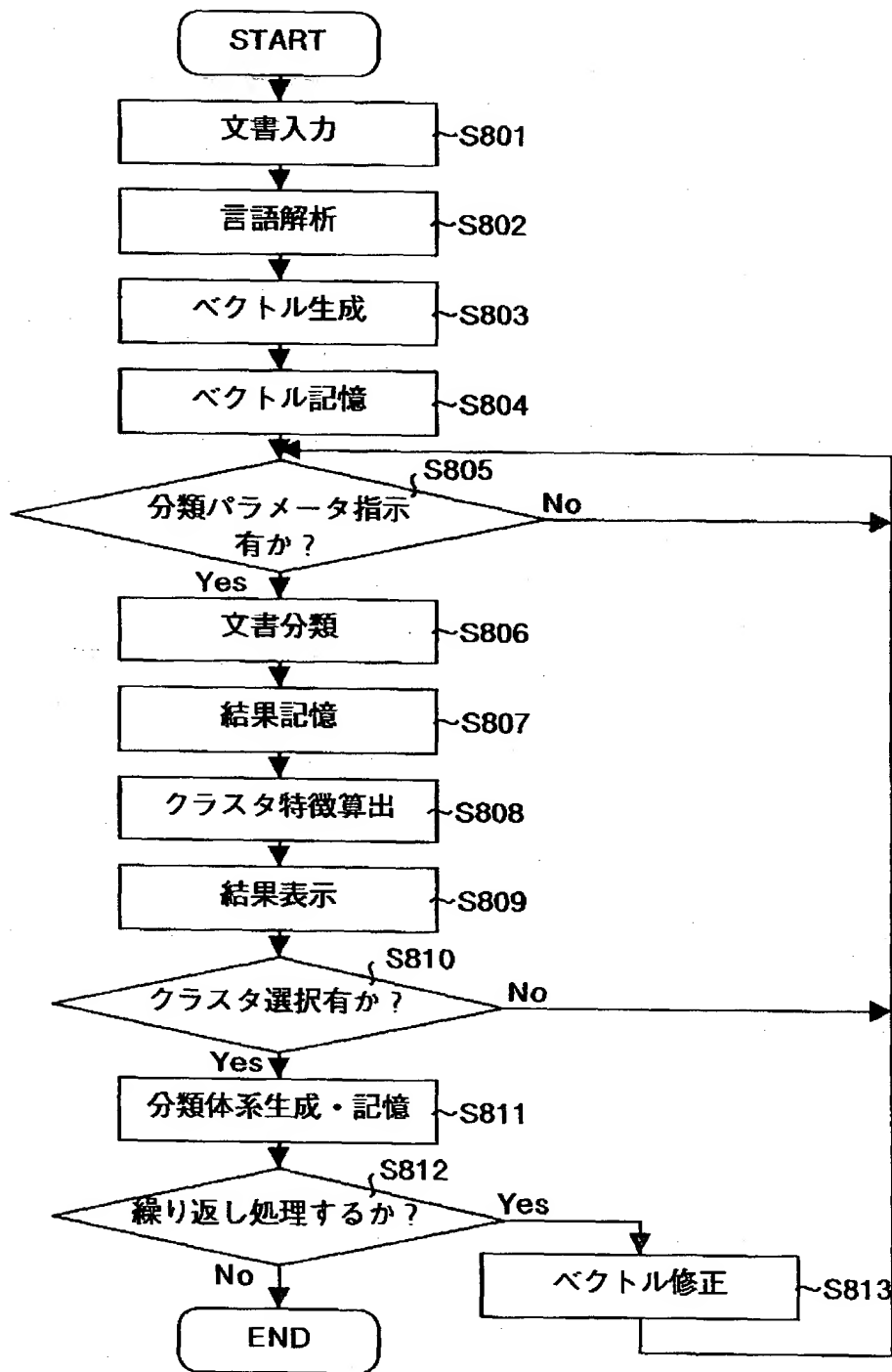
【図 6】



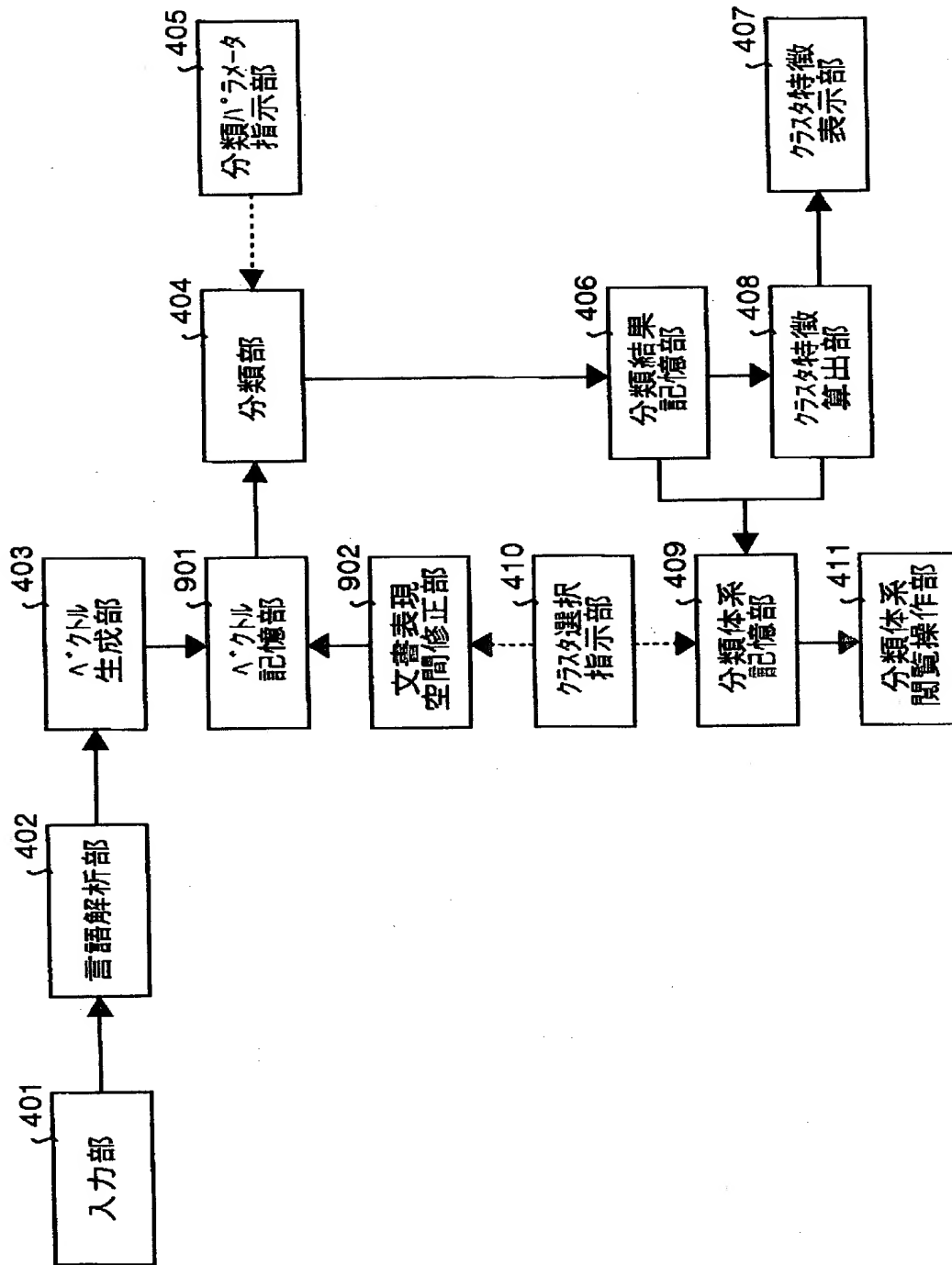
【図 7】



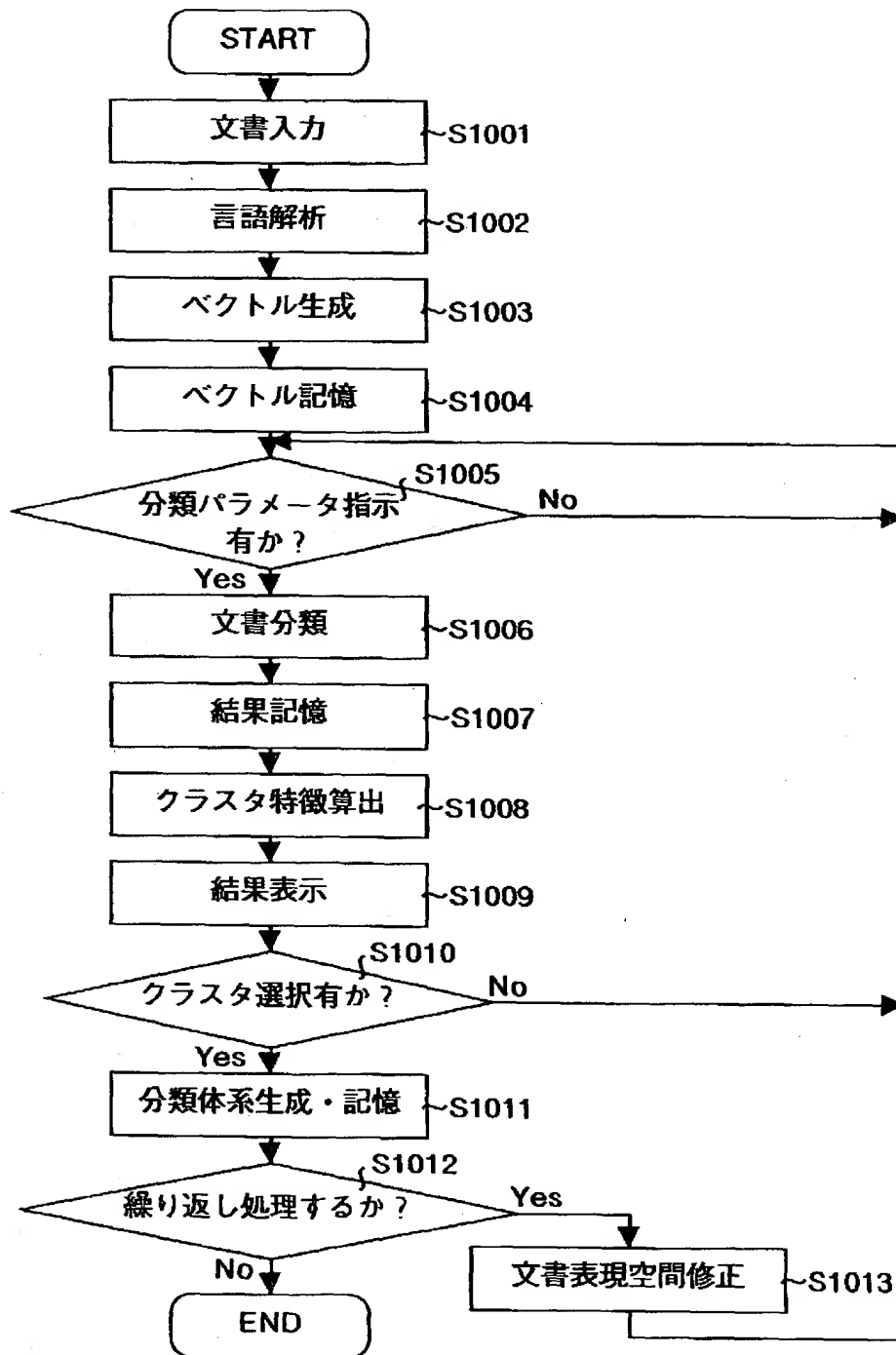
【図 8】



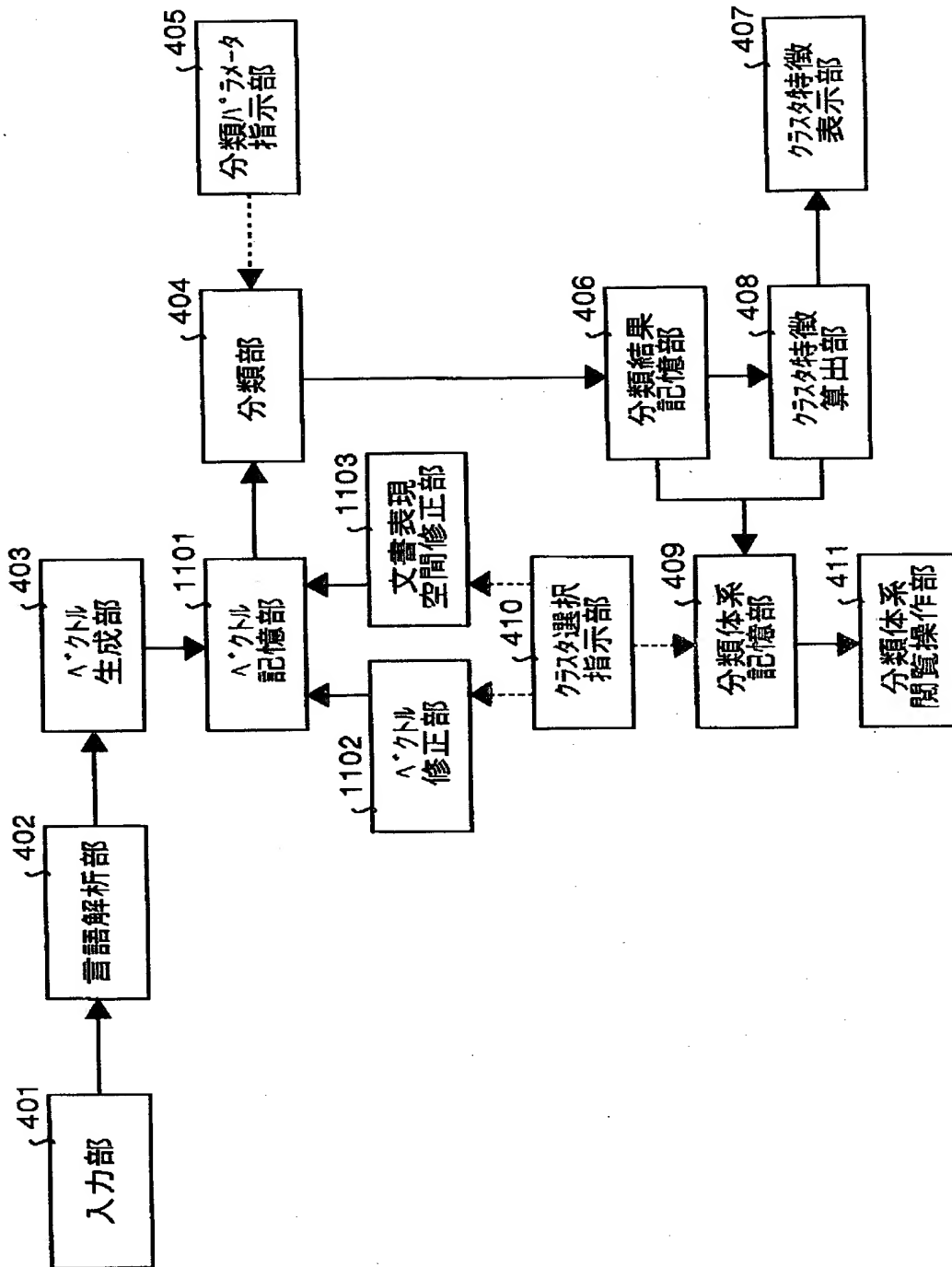
【図9】



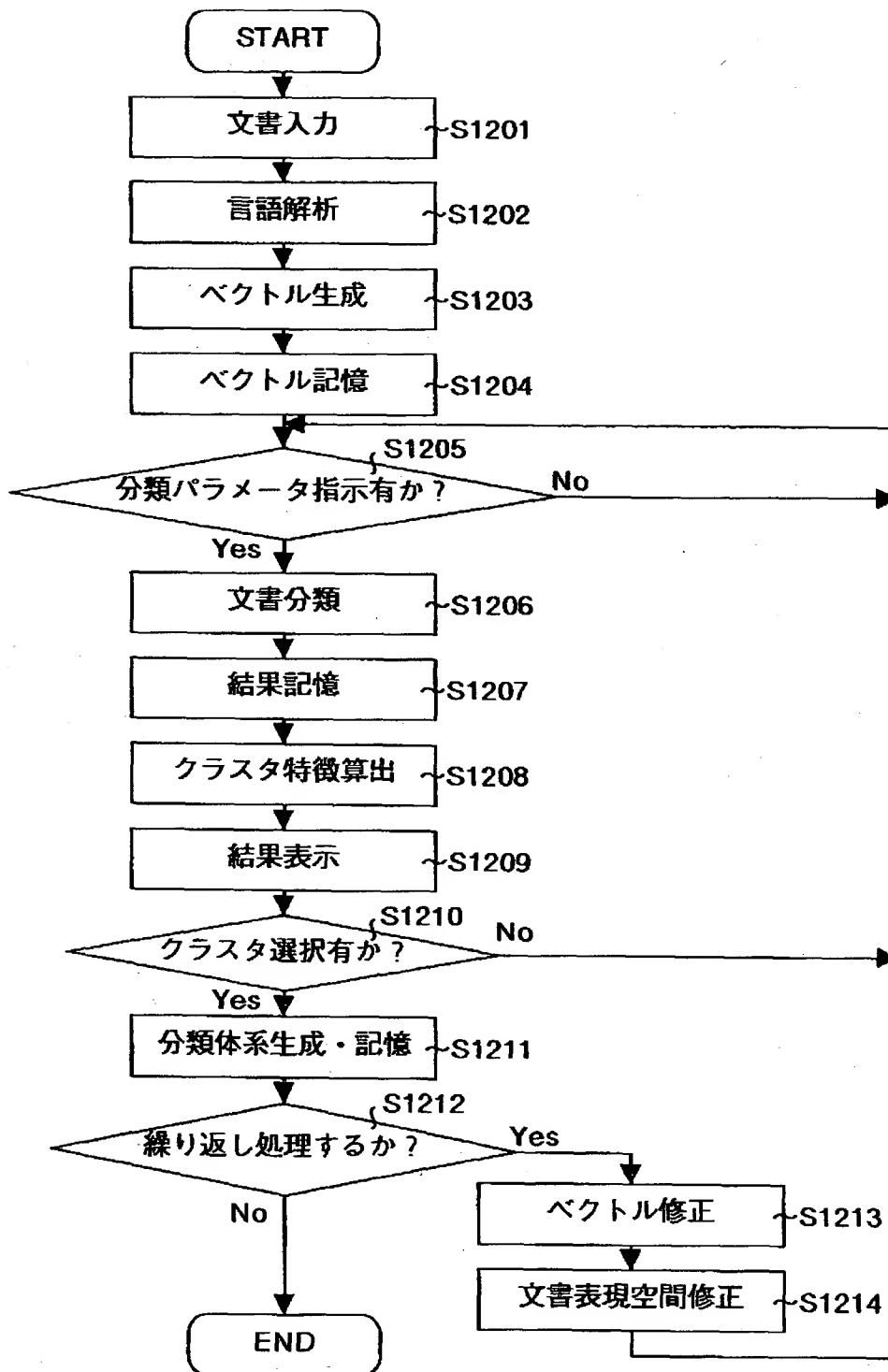
【図 10】



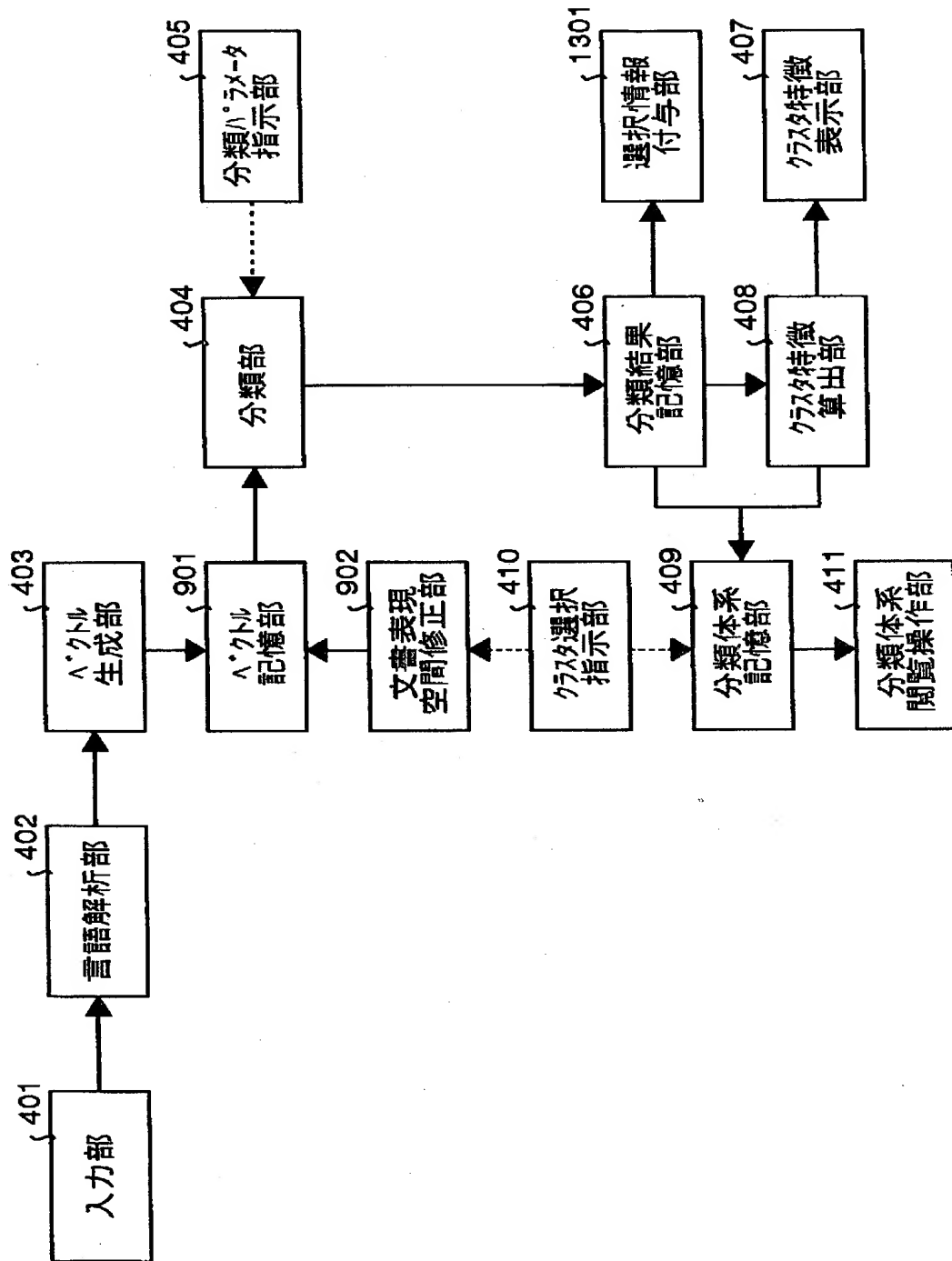
【図 11】



【図 12】



【図 13】

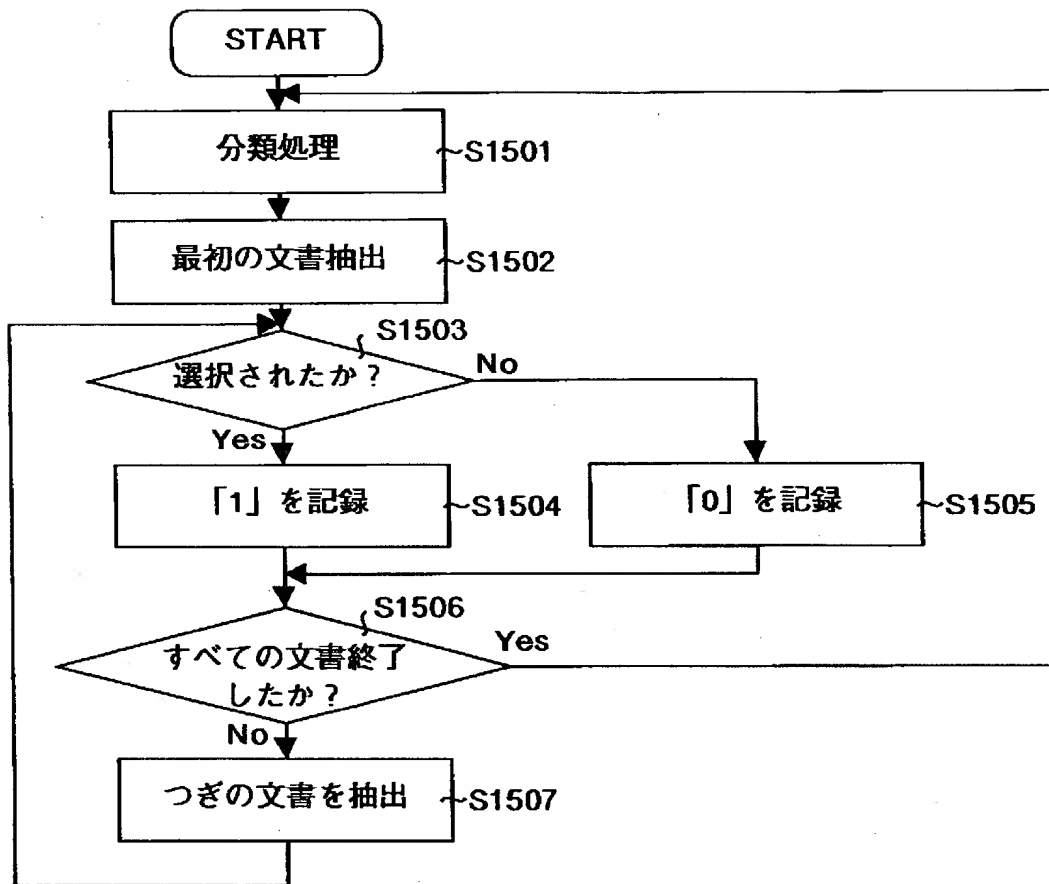


【図 1 4】

1400
S

文書ID	選択情報 (1 = 選択 / 0 = 未選択) 分類回数順
1	[1, 1, 0, 0]
2	[0, 0, 0, 0]
3	[0, 1, 0, 0]
⋮	⋮
n-1	[1, 0, 0, 0]
n	[0, 0, 0, 0]

【図 15】



【書類名】 要約書

【要約】

【課題】 文書の内容に基づいて文書の分類をおこなう文書分類装置において、任意の文書集合にどのような内容が含まれるかを漸次的に収集することを課題とする。

【解決手段】 文書データを入力する入力部401と、入力された文書データを解析して言語解析情報を得る言語解析部402と、得られた言語解析情報に基づいて文書データに対する文書特徴ベクトルを生成するベクトル生成部403と、生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類部404と、生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出部408と、算出されたクラスタ特徴を表示する表クラスタ特徴表示部407と、表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示部410と、選択されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶部409とを備える。

【選択図】 図4

出 願 人 履 歴 情 報

識別番号 [000006747]

1. 変更年月日 1990年 8月24日

[変更理由] 新規登録

住 所 東京都大田区中馬込1丁目3番6号

氏 名 株式会社リコー